Découverte de configurations de traits textuels pour la caractérisation des segments d'obsolescence

Marion Laignelet, Marie-Paule Péry-Woodley, Ludovic Tanguy

Découverte de configurations de traits textuels pour la caractérisation des segments d'obsolescence

Marion Laignelet* — **Marie-Paule Péry-Woodley*** — **Ludovic Tanguy***

* *CLLE-ERSS (CNRS - Université de toulouse)*
allées Antonio Machado – Toulouse
pery@univ-tlse2.fr, tanguy@univ-tlse2.fr

RÉSUMÉ. *Cet article présente une méthodologie de découverte de marqueurs envisagés comme des configurations de traits textuels pour la description et le repérage automatique de segments contenant des informations nécessitant des mises à jour (les segments d'obsolescence). La méthodologie mise en œuvre est fondée sur la prise en compte de traits textuels hétérogènes et à granularité variable. Nous mettons en place un système statistique à base de règles d'association pour faire émerger des données les combinaisons de traits pertinentes : traits intraphrastiques, hiérarchiques, positionnels et externes. Une évaluation de leur rôle en termes de performance est proposée. Nous travaillons sur un corpus de textes encyclopédiques annoté manuellement par des rédacteurs du monde de l'édition.*

ABSTRACT. *This paper presents a data-driven methodology for the automatic identification of text segments which contain information requiring updating ("obsolescence segments"). Our approach views markers as configurations of textual features and involves tagging text for a wide range of feature types of variable scope. We then apply a statistical method based on association rules whereby feature combinations relevant for the detection of obsolescence emerge from the data : intrasentential, hierarchical, positional and external features. We propose an evaluation of the respective roles of the different feature types. The study is based on a corpus of encyclopaedic texts which have been manually annotated by experts from the field of publishing.*

MOTS-CLÉS : *TAL, linguistique de corpus, discours, organisation textuelle, apprentissage automatique*

KEYWORDS: *NLP, corpus linguistics, discourse organisation, machine learning*

RSTI - DN – 13/2010. Le texte : analyse et exploration, pages 41 à 69

1. Introduction

La question de la mise à jour des documents se pose de façon cruciale dans les entreprises. Cette problématique est centrale pour les entreprises d'édition, notamment lorsqu'elles mettent à disposition leurs contenus éditoriaux sur internet. Ce travail s'inscrit dans une réalité professionnelle concrète : comment mettre à jour les articles encyclopédiques de la manière la plus exhaustive et le plus rapidement possible pour être en mesure d'offrir une information de qualité ? Nous proposons une réponse en abordant la question de la mise à jour des documents à travers la création d'un système d'aide au rédacteur d'articles encyclopédiques dont l'objectif est le repérage de segments textuels contenant de l'information obsolète ou de l'information susceptible d'évoluer dans le temps. Sur la base d'un repérage automatique de telles zones, c'est au rédacteur, *in fine*, de décider si l'information en question doit ou non être mise à jour et de quelle manière.

x. **Actualité**
§ Établir une liste exhaustive des **avancées récentes** de la recherche médicale est impossible tant les **progrès** sont nombreux. Toutefois, il convient de rappeler un certain nombre de **découvertes très récentes**. **En 2003,** l'une des grandes priorités de la recherche médicale internationale a concerné le sida.
x.y. Un vaccin contre le sida **?**
§ Des recherches portant sur les prostituées [...]. La recherche se tourne justement **aujourd'hui** vers des vaccins qui [...]. Des expériences ont été faites pour [...]. **En juin 2003,** une équipe de biologistes américains a obtenu des résultats qui **pourraient** laisser envisager [...]. Les chercheurs sont parvenus [...]. Cette découverte **pourrait** aboutir à la mise au point d'un antigène [...].

Source : Corpus ATLAS (fiche Médecine - Le Sida)

Exemple 1. *Un extrait d'encyclopédie à mettre à jour*

L'exemple 1 est un extrait des ressources encyclopédiques mises à disposition par la maison d'édition Atlas. L'auteur y exprime une issue possible et probable concernant les recherches sur le sida. La fiche dont est extrait ce passage a été publiée en 2003. Dans le cas d'une ré-édition, il serait bienvenu de vérifier et éventuellement mettre à jour ces informations. L'objectif de cette étude n'est ni de procéder à des calculs de la référence temporelle (Asher *et al.*, 1993), ni de chercher à associer un événement particulier à une date particulière comme c'est le cas en extraction d'information ou dans les systèmes de question-réponse. Nous ne cherchons pas non plus à valider la véracité d'événements particuliers. Nous recherchons les segments pour lesquels il est pertinent de penser que l'information donnée est susceptible d'avoir évolué entre le moment de l'édition de l'article et le moment de sa lecture. Dans l'exemple 1, la phrase « Cette découverte pourrait aboutir à la mise au point d'un antigène » doit être vérifiée et mise à jour si entre temps un antigène est découvert.

Un *segment d'obsolescence*[1] se définit comme une zone textuelle contenant des informations obsolètes ou évolutives qui devront faire l'objet d'une vérification (humaine) et éventuellement d'une mise à jour. Dans l'exemple 1, chacune des phrases constitue un segment d'obsolescence. La méthode de découverte de configurations de traits textuels pour caractériser les segments d'obsolescence est d'abord basée sur l'hypothèse que le caractère obsolescent ou évolutif d'une information peut se manifester concrètement dans les textes. Ainsi, les éléments en gras dans l'exemple 1 correspondent à des traits textuels potentiellement pertinents pour mettre au jour les segments d'obsolescence.

Plus précisément, notre méthodologie consiste à repérer des *traits textuels* variés sémantiquement et à granularité variable (Ho-Dac *et al.*, 2010). Nous adoptons une démarche très proche de celle développée par (Biber, 1988; Biber *et al.*, 2007) à laquelle nous ajoutons la prise en compte de traits structurels. Cet article porte sur le rôle des divers types de traits textuels pour le repérage de l'obsolescence : traits intraphrastiques, phrastiques, hiérarchiques (titres), positionnels et externes (domaine). Cinq étapes principales structurent notre méthode : tout d'abord, nous mettons en place une annotation manuelle du corpus en termes d'obsolescence ou non des zones textuelles ; deuxièmement, nous définissons un large ensemble de traits textuels susceptibles d'être de (bons) prédicteurs de l'obsolescence, tout en anticipant leur traitement automatique ; dans un troisième temps, nous projetons les traits textuels sur le corpus annoté manuellement ; la quatrième étape consiste à évaluer la corrélation entre les différents traits textuels et les segments d'obsolescence ; enfin, nous affinons les résultats de manière à définir, sur la base de ces traits, les marqueurs de l'obsolescence. Ces marqueurs sont utiles à deux niveaux : d'une part pour la description même des segments d'obsolescence, d'autre part parce qu'ils constituent le socle du système d'aide à la mise à jour des documents encyclopédiques destiné aux rédacteurs/auteurs des articles.

Dans la section 2, nous présentons le corpus de textes encyclopédiques sur lequel nous travaillons. Annoté manuellement, il permet à la fois de fournir une description des segments d'obsolescence et de mener un apprentissage automatique supervisé pour les caractériser finement. La section 3 décrit l'ensemble de traits textuels retenus qui semblent pertinents pour l'obsolescence ainsi que l'outil de traitement automatique des langues (TAL) construit pour les repérer automatiquement. Dans la section 4, nous proposons un modèle permettant de représenter la variabilité de grain des traits textuels et comment ils s'organisent entre eux. C'est une étape importante car elle permet de transformer des données textuelles en données appréhendables par des outils statistiques. Ces traitements ainsi que la méthode d'apprentissage automatique mise en œuvre sont développés dans la section 5. Enfin, le rôle des différents types de traits textuels est évalué dans la section 5.3.

1. Il s'agit d'un terme abrégé, utilisé pour « segment textuel contenant de l'information susceptible d'être obsolète ».

2. Une méthode en corpus

Pour comprendre et décrire l'obsolescence, nous nous basons sur un corpus de textes de type encyclopédique. L'utilisation d'un corpus permet :

– de nous assurer de la pertinence de l'objectif visé en n'utilisant que des textes issus du monde professionnel et nécessitant réellement une mise à jour régulière des informations contenues ;

– d'évaluer (quantitativement) la pertinence et la réalité du phénomène d'obsolescence dans ces textes ;

– de comprendre et décrire (qualitativement et à grande échelle) les segments d'obsolescence.

2.1. *Description du corpus*

Le corpus est composé de textes issus du monde de l'édition. Il est composé de deux sous-corpus : le sous-corpus *Atlas* qui contient des fiches encyclopédiques provenant des Éditions Atlas[2] et le sous-corpus *Larousse* qui a été constitué à partir d'articles extraits du *Grand Universel Larousse* (GUL) et du *Grand Larousse Informatisé* (GLI) et présentant des développements encyclopédiques moyens et longs.

Tous ces textes sont de type encyclopédique : ils ont une visée informative et potentiellement didactique sur des faits, événements, états, personnages, etc. Ils s'inscrivent dans des rubriques thématiques relativement consensuelles : histoire, géographie, arts et littératures, sciences et techniques, médecine, etc.

Le corpus est composé de 282 000 mots (soit environ 10 000 phrases). Les textes sont encodés en XML et nous avons conservé au maximum les informations de mise en forme disponibles (titres, niveaux de section, paragraphes, etc).

2.2. *Annotation manuelle : comprendre et évaluer l'obsolescence*

Une annotation manuelle des segments d'obsolescence est utile à deux niveaux : d'abord parce qu'elle permet de mieux comprendre le phénomène que nous souhaitons repérer automatiquement, également parce qu'elle permettra par la suite de mettre en place un apprentissage automatique (supervisé) des configurations de traits textuels pertinents dans les segments d'obsolescence.

L'annotation manuelle des segments d'obsolescence a été menée par quatre annotateurs : par l'une des auteurs de cet article[3] pour ce qui est de l'ensemble du corpus, et par des rédacteurs travaillant aux Éditions Larousse pour ce qui concerne le

2. Ce sont des éditions qui proposent des publications sous forme d'abonnement : le client reçoit un nombre déterminé de fiches encyclopédiques sur des domaines variés de façon régulière.
3. Marion Laignelet

sous-corpus *Larousse*. L'annotation manuelle effectuée par Marion Laignelet est ainsi confirmée par l'annotation des segments par les experts. La multi-annotation permet de mesurer le jugement d'obsolescence au-delà du sentiment subjectif individuel et donc de valider la réalité de cette notion. Le protocole d'annotation manuelle des segments d'obsolescence est relativement lâche afin d'évaluer la pertinence de la notion et son caractère intuitif : il était demandé aux rédacteurs de surligner les passages textuels du corpus qu'ils estiment être susceptibles d'être mis à jour dans l'immédiat ou à moyen terme.

Pour des raisons techniques, nous ramenons les segments d'obsolescence annotés à une taille uniforme correspondant aux frontières de la phrase. Lorsque l'information est locale, c'est-à-dire incluse dans une phrase, alors nous considérons la phrase entière comme un segment d'obsolescence. Lorsque l'annotation manuelle couvre plusieurs phrases, paragraphes ou sections, le segment est divisé en autant de sous-segments qu'il y a de phrases.

2.2.1. *Accord inter-juges*

Sur la base de ces annotations multiples, nous avons mesuré l'accord de jugement sur l'obsolescence à l'aide du *r* de Finn (Laignelet, 2009). Les scores sont situés entre 0,75 et 0,83 selon les annotateurs. Ces résultats expriment un consensus stable et objectif sur ce que les rédacteurs entendent par « segment d'obsolescence ». Ces résultats sont malgré tout assez loin de la valeur d'accord maximal (soit 1) ce qui permet de relativiser les performances du prototype visé.

En termes de proportion de segments obsolescents dans le corpus, 15,2 % des phrases du corpus sont jugées obsolescentes par les annotateurs. Une aide automatique au repérage de tels segments se justifie : il ne sont ni trop fréquents (ce qui impliquerait au final une relecture manuelle complète des documents), ni trop rares pour être ignorés.

L'annotation manuelle de l'obsolescence fait apparaître la diversité des segments d'obsolescence tant dans les types d'informations en cause que dans leur réalisation textuelle concrète. Nous présentons ci-après quelques exemples illustrant cette variété.

2.2.2. *Diversité au niveau informationnel*

Nous observons tout d'abord des cas où l'information ne semble plus pertinente parce que du temps s'est écoulé depuis sa rédaction.

Dans l'exemple 2, le rédacteur informe son lecteur que le nombre de nouveaux cas de maladies du travail est de 160 millions en 2001[4]. C'est une donnée qui, malgré l'inexorabilité temporelle, restera toujours vraie. Cependant, en tant que lecteur, il semble plus approprié de disposer de données plus proches de l'année de la lecture[5]. Ce qu'une encyclopédie est censée apporter, du moins pour ce qui est du domaine géo-

4. L'article a été écrit en 2003.
5. Soit 2006 au moment où le corpus est constitué et annoté.

graphique, c'est un état des faits de la société le plus réel et le plus actuel possible pour un phénomène donné. Pour les articles relevant du domaine de l'histoire, la demande est différente.

> L'organisation mondiale de la santé (OMS) estime, en effet, à 160 millions le nombre de nouveaux cas dans le monde, en 2001.
>
> Source : ATLAS (fiche Médecine - Les maladies professionnelles)

Exemple 2. *Un segment d'obsolescence contenant une information non actuelle*

Dans d'autres cas, le rédacteur formule des prédictions à court terme et/ou précise temporellement (exemple 3) ou à plus long terme (exemple 4).

Dans l'exemple 3, l'auteur émet des hypothèses sur la production de pétrole pour l'année 2004. Or, si l'on considère le repère temporel de constitution et d'annotation du corpus, soit l'année 2006, la prédiction date de 2 ans : cette phrase entière est donc devenue obsolète.

> La production en off-shore est en plein essor : elle devrait atteindre le million de barils par jour en 2004.
>
> Source : ATLAS (fiche Afrique - L'afrique australe)

Exemple 3. *Un segment d'obsolescence contenant une information nécessairement erronée*

L'exemple 4 rend compte d'un cas où l'auteur fait des suppositions quant à la possibilité d'effectuer des greffes. Il n'y a pas de date mais le conditionnel « devrait » permet d'interpréter l'événement comme inaccompli.

> De même, selon une équipe de chercheurs suédois, il devrait être possible d'ici deux ou trois ans d'effectuer des greffes d'utérus.
>
> Source : ATLAS (fiche Médecine - La recherche médicale)

Exemple 4. *Un segment d'obsolescence contenant une information à vérifier*

2.2.3. *Diversité au niveau textuel*

La complexité des segments d'obsolescence se manifeste également sur le plan textuel. L'information nécessitant une mise à jour peut être incluse dans des segments locaux (noms, syntagmes) ou des segments de plus grande taille (phrase, paragraphe, section). Dans l'exemple 5, l'annotation manuelle porte uniquement sur la seconde partie de la proposition, soit « est tombé à 11 ‰ en 2003 ».

Dans d'autres cas, le segment couvre plusieurs propositions, voire le paragraphe ou la section entière. Dans l'exemple 6, les deux dernières propositions ont été annotées

Le taux de natalité, encore situé entre 20 et 25 ‰ jusqu'aux années 1960, est tombé à 11 ‰ en 2003.

Source : Corpus GUL (fiche Géographie - Géorgie)

Exemple 5. *Un segment d'obsolescence local*

x. Le travail, c'est la santé ?
Ainsi, en France, les chiffres officiels avancés par les autorités sanitaires ne reflètent pas la réalité car ils concernent uniquement les cas indemnisés. Pour autant, la situation s'améliore et on note une augmentation régulière des cas déclarés depuis le début des années 1990.

Source : ATLAS (fiche Médecine - Les maladies professionnelles

Exemple 6. *Un segment d'obsolescence à gros grain*

comme un segment d'obsolescence. Ce segment sera, comme cela a été expliqué à la section 2.2, redécoupé en phrases.

La section 3 décrit plus précisément les traits textuels qui nous semblent pertinents pour décrire l'obsolescence ainsi que la méthode pour les marquer automatiquement dans le corpus.

3. Linguistique, discours et TAL

Nous nous inscrivons dans le cadre de l'analyse de l'organisation textuelle qui considère que les textes ne sont pas simplement des suites de phrases ou de mots. Nous pensons, à la suite de (Charolles, 1995; Péry-Woodley, 2005) ou encore (Ho-Dac *et al.*, 2004) que le texte est une unité complexe qui suit un principe central de cohérence. Deux types d'objectifs sont généralement visés : repérer et décrire des blocs homogènes et mettre en lumière des relations entre ces blocs.

Dans ce cadre théorique, nous cherchons à décrire les segments textuels sur la base de la présence de traits textuels spécifiques qui, parce qu'ils co-occurrent dans les textes, vont permettre une interprétation particulière d'un segment textuel donné, pour nous l'obsolescence.

Un *trait textuel* fonctionne ici comme une sonde projetée sur le texte. La méthode que nous employons, en utilisant des données annotées et des procédures statistiques, permettra de décider dans quelle mesure il est lié au phénomène visé. Nous considérons qu'un *marqueur* de l'obsolescence peut être réalisé par :

– un trait univoque : par exemple, une expression temporelle référant à une période située dans le futur a de grandes chances d'introduire un segment obsolescent ;

– une combinaison de traits différents : par exemple, un conditionnel associé à une expression de mesure situé dans une phrase en position de conclusion introduit souvent une prédiction ou une supposition de la part de l'auteur.

D'autres travaux s'intéressent à la question des marqueurs linguistiques composés d'éléments hétérogènes et à granularité variable (Ho-Dac *et al.*, 2010). Cette méthode a la particularité de déployer un grand nombre de traits textuels, de différents types et sans nécessairement cibler finement *a priori* les traits précis et fiables.

3.1. *Choix des traits textuels*

Le choix des traits textuels susceptibles d'être de bons prédicteurs de l'obsolescence s'est fait à la fois à partir de l'observation du corpus annoté et de travaux antérieurs. Nous nous sommes notamment inspirés des travaux de (Borillo, 1997; Desclés *et al.*, 1992; Gosselin, 2005; Weinrich, 1973; Laurendeau, 2004; Gosselin, 2005) sur le temps, l'aspect, la modalité, de ceux de (Power *et al.*, 2003; Teufel, 1999; Bouffier, 2008) pour ce qui est de la prise en compte de la structure des documents ou encore des travaux de (Charolles, 1997) sur les introducteurs de cadre, de (Jacques *et al.*, 2006) sur les titres, de (Mani, 2001) enfin pour ce que est de l'importance de la position des traits dans la phrase, dans le paragraphe, dans la section.

Nous proposons une méthodologie pour mettre en lumière des traits et/ou des combinaisons de traits susceptibles de favoriser ou non l'interprétation obsolescente d'un segment textuel. Pour atteindre cet objectif, nous avons mis en évidence une vaste palette de traits textuels potentiellement pertinents pour l'obsolescence que nous repérons et typons de manière automatique.

3.2. *Repérage et typage automatiques des traits textuels susceptibles de marquer l'obsolescence*

Nous présentons dans cette section un outil développé dans le but de repérer et de typer les expressions correspondant à chacun des types de traits textuels susceptibles d'être pertinents pour la description et le repérage des segments d'obsolescence.

Ce processus est entièrement automatique et a été développé à l'aide de la plateforme LinguaStream[6] (Widlöcher *et al.*, 2005). Il s'agit d'un environnement de développement dédié aux traitements TAL qui facilite la création de segmenteurs (en mots, en phrases), de lexiques, de grammaires prolog, de macro expressions régulières.

Pour notre étude, nous avons développé un ensemble de traitements spécifiques réunis dans l'outil ALIDIS[7] (Laignelet, 2009). Cet outil regroupe plusieurs modules

6. http ://www.linguastream.org/whitepaper.html
7. http ://marion.laignelet.free.fr

de repérage et d'annotation sémantique : modules de traitement du temps, de la modalité, de l'aspect, des positions, etc.

3.2.1. *Les traits intraphrastiques*

Les traits intraphrastiques sont des traits qui sont inclus entre les frontières des phrases. Ils sont sémantiquement assez variés.

L'analyseur temporel est chargé de repérer et de typer les expressions temporelles. Cela concerne les adverbiaux temporels et les temps verbaux.

Les adverbiaux temporels sont typés selon deux caractéristiques :

– leur nature : ce trait peut prendre les valeurs suivantes : *ponctuel, inachevé, déictique, durée, itération* ;

– la situation temporelle, qui peut prendre les valeurs suivantes : *antériorité++, antériorité, coïncidence, postériorite, indéterminé*[8].

Voici quelques exemples :

(1) pendant dix ans $\begin{bmatrix} nature : durée \\ sitTps : indéterminé \end{bmatrix}$

(2) en 1930 $\begin{bmatrix} nature : ponctuel \\ sitTps : antériorité++ \end{bmatrix}$

(3) depuis 2005 $\begin{bmatrix} nature : inachevée \\ sitTps : coïncidence \end{bmatrix}$

(4) aujourd'hui $\begin{bmatrix} nature : déictique \\ sitTps : coïncidence \end{bmatrix}$

Les temps verbaux sont directement issus des résultats de l'analyseur morpho-syntaxique TreeTagger[9]. Le typage des verbes est le suivant : le trait `temps` peut prendre les valeurs suivantes : *passé-composé, passé antérieur, plus-que-parfait, futur antérieur, conditionnel passé, présent, passé simple, imparfait, futur, conditionnel*.

L'analyseur de périphrases verbales permet de typer certaines expressions selon qu'elles expriment une action débutant, en cours ou achevée.

L'exemple suivant illustre le type d'expression repéré et annoté par cet analyseur :

8. Nous découpons le passé de manière arbitraire : *antériorité++* correspond aux dates antérieures à 1950, *antériorité* aux dates comprises entre 1951 et 1990, *coïncidence* aux dates comprises entre 1991 et 2007 et enfin *postériorite* fait référence aux dates situées après 2008 sur l'échelle du temps. Ce découpage est issu de notre analyse des besoins en termes de mise à jour de l'information.
9. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

(5) Les 3e et 4e tranches [...] [sont en cours de] [$accomplissement : déroulement$] réalisation.

Un analyseur d'entités nommées repère et type des expressions variées. Nous nous intéressons à huit classes principales, certaines étant susceptibles de se subdiviser en sous-classes. Ainsi une expression peut être classée comme un *lieu*, classe qui se subdivise en sous-classes de type *rivière, pays, ville, montagne*. Par exemple :

(6) Les expéditions d'Arzila [à Tanger] $\left[\begin{array}{l} classe : lieu \\ sousClasse : ville \end{array}\right]$

Cet analyseur repère également des expressions relevant des classes *personne*, *sigle*, *web*, *mail*, *marque*, *géopolitique*, *mesure* ou encore *superlatif*.

(7) La majeure partie du butin ramené par [Drake] [$classe : personne$]

(8) Le trafic de la [VOC] [$classe : sigle$] est centré sur le commerce du poivre et d'autres épices.

(9) les litiges supérieurs à [7 600 euros] $\left[\begin{array}{l} classe : mesure \\ sousClasse : évolutif \end{array}\right]$

(10) Le [taux de chômage] [$classe : géopolitique$] s'est effectivement effondré.

(11) dans [les zones les plus peuplées] $\left[\begin{array}{l} classe : superlatif \\ sousClasse : plus \end{array}\right]$ comme le Bas-Congo

L'analyseur d'expressions du point de vue donne des informations sur le positionnement de l'auteur vis-à-vis des propos qu'il écrit. Nous proposons neuf types différents : *distance, jugement, récence, prévision, importance, jugement personnel, source, thématique, restriction, argumentation, superlatif*.

Le type *jugement* donne des informations modalisantes sur le point de vue de l'auteur :

(12) La prolifération des États signataires de la Charte des Nations unies renforce [peut-être] [$type : jugement$] l'[impression] [$type : jugement$] d'homogénéité juridique de la communauté internationale.

Les types *récence* et *prévision* sont orientés temporellement. Il s'agit de syntagmes nominaux contenant un adjectif temporel qui ne peut être interprété que si la date à laquelle l'article a été écrit est connue :

(13) le [dernier Mondial] [$type : récence$] s'est tenu à

(14) les [recherches à venir] [$type : prévision$]

Le rédacteur peut également insister sur l'importance d'un fait à un moment donné :

(15) Il s'agit d'[un véritable enjeu] [$type : importance$]

Il peut également se distancier des propos qu'il avance en citant ses sources :

(16) on estime [$type : distance$]

(17) Selon le rapport de l'INSEE [$type : source$]

L'introduction de définitions introduit une certaine volonté de clarification des propos :

(18) On distingue [$type : définition$] deux classes :

Enfin, nous prenons en compte les organisateurs de l'argumentation car ils permettent au rédacteur de structurer ses propos. Ainsi, nous repérons des éléments argumentatifs comme « d'abord », « puis », « dans un premier temps », ainsi que des expressions comme « À ce sujet/propos » ou encore « Pour ce qui est de la dette ».

3.2.2. *Les traits phrastiques*

Un analyseur est chargé de typer les phrases selon qu'elles expriment une exclamation, une assertion ou une interrogation. Cet analyseur est basé sur le repérage de la ponctuation essentiellement.

3.2.3. *Les traits hiérarchiques*

Nous supposons que les titres ont un rôle à jouer dans la description de l'obsolescence. Par exemple, dans l'exemple 1 donné en introduction, le fait que le mot « actualités » se trouve dans le titre oriente l'ensemble de la section qui suit. Les phrases de la section héritent ainsi, sous la forme de traits hiérarchiques, les traits intraphrastiques du titre qui les couvre. Dans cet exemple, *actualité* génère un trait de type *point de vue* (*récence*), qui va se propager à toutes les phrases de la section.

3.2.4. *Les traits positionnels*

Trois analyseurs ont pour but le marquage de positions. Tout d'abord, nous exploitons la position des traits textuels présentés jusque-là dans la phrase : un trait peut alors être situé en position initiale ou finale de la phrase ou en position d'amorce. L'amorce correspond aux cas où une expression est immédiatement suivie de deux-points et introduit une définition, une explication, un résumé. Ce type de présentation

est très fréquent dans les fiches encyclopédiques de la société Atlas qui met en avant une mise en forme des informations riche et complexe[10].

(19) La population : le nombre d'habitants ne cesse d'augmenter.

Un autre analyseur informe sur la position de la phrase dans le paragraphe qui peut alors prendre les valeurs suivantes : *début de paragraphe, fin de paragraphe, phrase seule dans le paragraphe*. Le troisième analyseur évalue la position du paragraphe dans le document et attribue les traits suivants : *début de zone, fin de zone* pour ce qui est des sections de niveau 1 et *début de division, fin de division* pour ce qui est des sections de niveau 2.

3.2.5. *Les traits externes*

Cet analyseur récupère dans les métadonnées l'information concernant le domaine dans lequel s'inscrit l'article et qui est par ailleurs fournie par l'éditeur. Nous avons harmonisé les différents domaines en 11 catégories : géographie, histoire, sciences et techniques, arts et littératures, médecine, économie, société, droit, biologie, sport, divers.

3.3. *Évaluation des traits intraphrastiques*

Le tableau 1 indique, indépendamment de la notion d'obsolescence, le nombre total d'occurrences des traits linguistiques repérés de manière automatique dans notre corpus, ainsi qu'une évaluation du rappel et de la précision des analyseurs correspondants. Cette évaluation a été menée à la main sur $1/10^e$ du corpus annoté automatiquement. L'évaluation porte à la fois sur la performance de repérage et celle de typage.

	Nombre d'occurrences	Précision	Rappel
Temps verbaux	15 768	97 %	98 %
Adverbiaux temporels	4 459	92 %	98 %
Périphrases verbales	85	99 %	43 %
Entités nommées	12 306	99 %	83 %
Expression du point de vue	916	73 %	98 %
Moyenne	33 534	93 %	85 %

Tableau 1. *Performance globale de l'outil* ALIDIS

On observe quelques disparités. Tout d'abord, le repérage des périphrases verbales (« des recherches sont en cours », « les essais sont terminés ») a une précision correcte mais un rappel médiocre (43 %) : le repérage de ce type d'expression nécessiterait un cataloguage approfondi. Dans une moindre mesure, la situation est identique pour le

10. Pour des exemples de visualisation voir (Laignelet, 2009, p. 37-41).

repérage des entités nommées[11]. Concernant les expressions du point de vue, le rappel est correct mais la précision est moyenne (73 %) : le repérage automatique de telles expressions est plus délicat notamment parce qu'il nécessite en lui-même une prise en compte plus large du contexte. Par exemple, nous souhaitons repérer l'expression « la recherche [prévoit] des avancées considérables dans ce domaine » mais pas une expression comme « la loi [prévoit] un an de prison pour... » ; or le fait que nous nous basons sur la présence du verbe *prévoir* nous renvoie les deux cas. Il faut ajouter à ces comptages, 17 830 occurrences de traits de type positionnels et externes, ce qui nous amène à traiter 51 364 occurrences de traits, tous types confondus.

Les performances globales de l'outil ALIDIS sont acceptables : un rappel de 85 % et une précision de 93 % en moyenne. Ces résultats nous permettent d'envisager une exploitation automatique et à grande échelle.

3.4. *Exemple*

Les figures 1 et 2 montrent comment et avec quels traits textuels l'exemple 1, présenté en introduction, a été annoté. La première figure illustre le cas des traits intraphrastiques ; la seconde, le cas des traits structurels (phrastiques, hiérarchiques, positionnels).

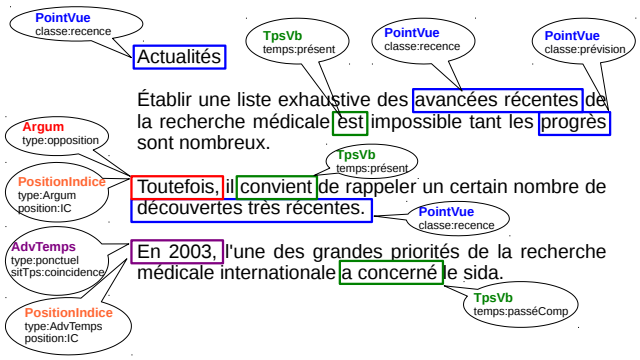

Figure 1. *Repérage et annotation automatique des traits intraphrastiques*

Le corpus initialement annoté manuellement au niveau des segments d'obsolescence est enrichi d'une projection automatique des traits textuels. Des traitements sta-

11. On compte plus précisément : 37,7 % d'entités nommées de type *lieu*, 12,8 % de type *personne*, 8,8 % de type *sigle*, 15,8 % de type *mesure*, 20,6 % de type *géopolitique* et 4,3 % de type *superlatif*.

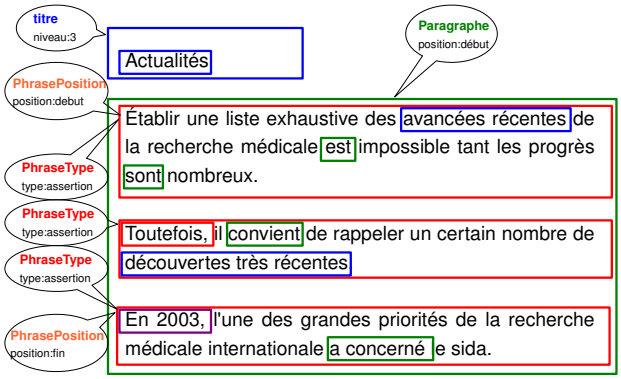

Figure 2. *Repérage et annotation automatique des traits structurels : titres, phrases et positions*

tistiques sont ensuite mis en place afin de mettre au jour des combinaisons de traits pertinentes pour l'obsolescence. Mais il est nécessaire de transformer les repérages et annotations en un format exploitable par des outils statistiques. Pour cela, nous proposons une étape intermédiaire de modélisation des données.

4. Un modèle de représentation des traits textuels à granularité variable

Pour pouvoir traiter statistiquement l'ensemble des traits textuels marqués de manière automatique dans le corpus, il faut passer d'un format textuel à une matrice permettant d'exploiter les données avec des outils statistiques. Nous souhaitons notamment pouvoir représenter les informations textuelles dans un tableau d'individus (dans notre cas, les phrases) caractérisés par un ensemble de variables (dans notre cas, les traits textuels). La principale difficulté réside dans la gestion des différences de granularité des traits textuels pris en compte et de leur imbrication.

Pour arriver à cette transformation, nous proposons un modèle de représentation des traits textuels permettant de rendre compte de leur organisation. Il est conçu pour suivre les quatre principes de représentation présentés dans le tableau 2.

Le principe 2 est lié au fait que dans notre corpus, les titres ne nécessitent pas de mise à jour[12]. Par ailleurs, nous craignons une redondance informationnelle si les traits

12. Aucun titre n'a été annoté manuellement par les experts.

textuels présents dans les titres sont traités à la fois comme segments obsolescents et aussi comme segments prédicteurs de l'obsolescence. Cela fausserait sans doute les résultats.

Le principe 3 permet de rendre l'outil évolutif : nous souhaitons un système facilitant l'ajout, la suppression ou la modification des traits textuels et/ou de leur typage sans avoir à tout réécrire à chaque modification. Le nombre et la nature des traits (variables) décrivant chaque phrase est donc calculé dynamiquement en fonction des données d'entrée.

Principe 1	Un titre et une phrase doivent pouvoir être décrits à travers n'importe quel type de trait : intra-phrastique, hiérarchique, positionnel phrastique et textuel, externe
Principe 2	Une phrase peut être obsolescente mais un titre ne le sera jamais, il est un prédicteur potentiel de l'obsolescence (cf. la notion d'héritage de contexte (Zerida *et al.*, 2006))
Principe 3	On ne connait pas *a priori* le nombre de traits présents à l'intérieur de la phrase et du titre ni le niveau de profondeur du segment ; de plus, on veut pouvoir modifier les traits et leur typage sans avoir à réécrire tous les programmes.
Principe 4	On ne connait pas *a priori* le format d'entrée pour les statistiques ; le stockage des données doit être au plus près de la réalité des textes (en termes de relations entre les traits et les segments que l'on souhaite décrire).

Tableau 2. *Les principes de représentation*

Le schéma UML de la figure 3 permet de représenter les unités discursives définies comme élémentaires pour nos besoins ainsi que les relations qu'elles peuvent entretenir entre elles. Nous avons choisi d'utiliser le langage UML[13] parce qu'il fournit un cadre standart strict et une définition précise des outils de représentation.

D'une manière générale, tout élément linguistique annoté est considéré comme une unité discursive. À chaque unité discursive est associée une structure de traits apportant des informations de type sémantique, syntaxique, morphologique, etc. Cette unité discursive peut être réalisée par n'importe quel élément linguistique qui a été annoté soit par l'outil ALIDIS, soit manuellement.

Au sein de ce modèle, dont la portée se veut plus générale que celle de l'étude, nous distinguons deux grandes catégories d'unités discursives : les unités d'analyse et les indices discursifs. Le choix des unités d'analyse est fait par le linguiste analyste. Dans notre cas, il s'agit des phrases et les titres. Ces unités peuvent avoir des rôles différents dans les textes : il est nécessaire de pouvoir les traiter de manière différente, certaines unités étant régies, d'autres rectrices. Ce choix permet de traiter les phrases comme des unités régies par des titres qui hériteront des traits textuels des unités rectrices, les titres en l'occurrence.

Les indices discursifs sont réalisés par les autres unités présentes dans le texte. Ils peuvent être de deux types. Les indices englobés font référence aux traits textuels inclus dans l'unité d'analyse en question : dans notre cas, il s'agit des éléments intraphrastiques essentiellement. Les indices englobants sont les traits textuels de taille plus grande que l'unité d'analyse : dans notre cas, il s'agit des éléments positionnels

13. www.uml.org

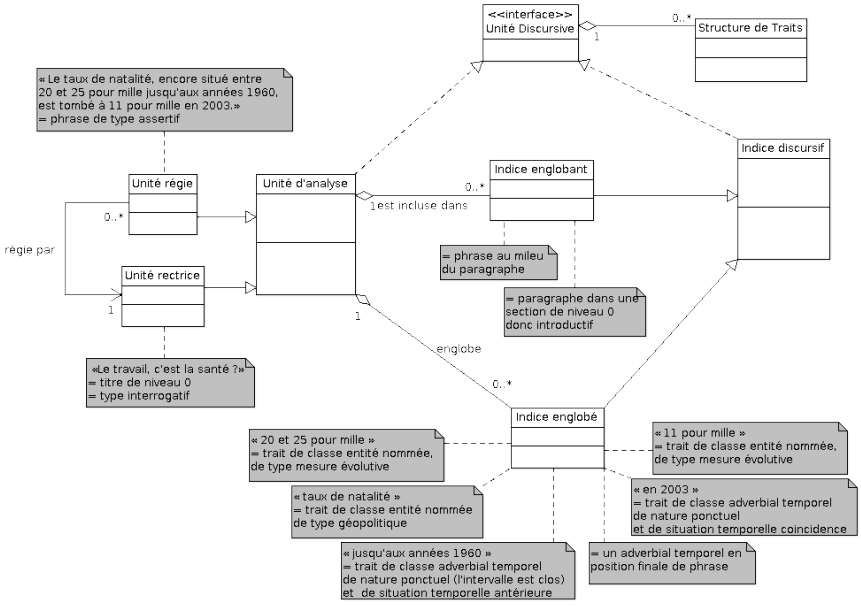

Figure 3. *Modèle UML permettant de gérer la variabilité du grain d'analyse des traits textuels*

et externes. Une unité d'analyse peut être associée à *n* indices englobants et *n* indices englobés.

Les limites principales de ce modèle sont les suivantes :

– la non-représentation des relations linéaires entre unités de même classe (notamment entre indices englobés) ;

– la position des expressions linguistiques dans la phrase est aujourd'hui mal gérée. La position est un trait qui s'ajoute à une expression linguistique (elle aussi un trait) : il y a donc redondance et de fait un biais interprétatif possible. Il serait possible d'améliorer ce point en utilisant un analyseur capable de refermer les cadres de discours ouverts par les adverbiaux à l'initiale de la proposition (*i.e.* les introducteurs de cadres (Charolles, 1997; Bilhaut, 2006)). L'adverbial en position initiale serait alors traité comme les titres, soit comme une unité rectrice de l'ensemble des phrases qu'il englobe.

Cette étape de modélisation nous permet de penser les éléments linguistiques sur un même plan tout en gardant les informations propres à chacun. Cette modélisation constitue le pivot permettant de transformer et d'organiser les annotations manuelles

de l'obsolescence et les traits textuels au sein d'une base de données relationnelle (décrite dans (Laignelet, 2009)). De cette base de données sont finalement extraites les informations représentées en deux dimensions : des variables (les traits textuels) caractérisent des individus (les phrases).

La suite de ce travail consiste à valider la pertinence des traits textuels considérés en système : quels traits et quelles combinaisons de traits sont corrélés à la variable *obsolescence* ? En d'autres termes, quels traits textuels sont à même de devenir des marqueurs de l'obsolescence ?

5. Description et analyse statistique

Maintenant que nous disposons d'une représentation des unités de discours sous un format matriciel, nous pouvons mener une série de tests statistiques afin de *(i)* comprendre et mesurer l'organisation des traits textuels dans les segments d'obsolescence annotés manuellement et *(ii)* mesurer l'impact des éléments linguistiques hiérarchiques et positionnels pour notre tâche.

Le choix de passer par les statistiques nous semble évident car il s'agit d'un outil approprié pour :

– traiter des données conséquentes en termes de quantité et de diversité des variables ;

– confirmer/infirmer des hypothèses de recherche ;

– faire émerger des informations nouvelles.

Nous avons mené à la fois des analyses statistiques de type descriptif et des analyses statistiques prédictives. Dans les deux cas, l'objectif est de décrire les segments d'obsolescence ainsi que les traits textuels fonctionnant en configuration. Dans le second cas, cela permet de capitaliser des connaissances nouvelles pour les réutiliser sur de nouvelles données.

Nous travaillons maintenant sur les données transformées en matrice : celle-ci est constituée de 9 916 individus[14] caractérisés par 150 variables[15].

Le format de base de données permet de créer facilement des vues différentes selon les objectifs visés. Dans un premier temps, deux vues différentes sont créées : une vue comprenant des données quantitatives (indication du nombre de verbes au présent dans une phrase par exemple) pour les statistiques descriptives et une vue où les valeurs des variables sont ramenées à 0/1, soit présence/absence, pour la technique d'apprentissage automatique que nous utilisons (cf. les règles d'association, section 5.2.1).

14. Chaque individu est relié à une phrase du corpus initial décrit dans la section 2.
15. Soit l'ensemble de tous les traits textuels décrits dans la section 3.2.

5.1. *Statistiques descriptives*

5.1.1. *Statistiques de base*

Afin de vérifier la pertinence des traits textuels susceptibles de jouer un rôle dans les segments d'obsolescence, nous avons mesuré dans un premier temps la corrélation de la variable *obsol* (interprétée à partir des annotations manuelles de l'obsolescence par les experts) avec chacune des autres variables de la base (les traits repérés de manière automatique par l'outil ALIDIS). Nous avons utilisé le logiciel SPAD[16]. Cette mesure de la pertinence des traits met au jour :

– des traits corrélés positivement[17] à la variable *obsol*, c'est-à-dire des traits qui apparaissent préférentiellement dans les segments d'obsolescence : par exemple, les entités nommées de type *géopolitique* (« nombre d'habitants », « PIB » ; v-test à 26,00) ou de type *mesure évolutive* (« 30 hab/km2 » ; v-test à 24,39) ou encore les adverbiaux temporels de type *déictique coïncidence* (« aujourd'hui » ; v-test à 18,91),

– des traits corrélés négativement à la variable *obsol*, c'est-à-dire des traits qui apparaissent moins fréquemment dans les segments non obsolescents : par exemple, les entités nommées de type *personne* (« Mohandas Karamchand Gandhi » ; v-test à -5,95) ou encore les adverbiaux temporels de type *ponctuel antériorité++* (« en 1800 » ; v-test à -6,18) ,

– des traits non corrélés à la variable *obsol*, c'est-à-dire des traits qui apparaissent indifféremment dans les segments obsolescents et dans les segments non obsolescents : par exemple, le futur (v-test à 0,72).

Ces premiers calculs indiquent tout d'abord que les traits linguistiques pris en compte se justifient. En effet, sur les 150 variables étudiées, 53 sont corrélées positivement, 21 négativement et 76 sont neutres. Parce que nous recherchons à terme des combinaisons de traits textuels, nous faisons le choix de conserver pour les calculs qui vont suivre l'ensemble des traits, y compris ceux dont la corrélation avec la variable *obsol* est nulle. En effet, nous supposons qu'un trait, alors qu'il est insignifiant pour l'obsolescence lorsqu'il est seul dans un segment, est capable d'orienter l'interprétation obsolescente d'un segment s'il cooccurre avec un ou plusieurs autres traits.

Les corrélations mises à jour avec cette méthode mettent en avant des tendances intéressantes. Par exemple, les adverbiaux temporels de type *déictique* (« aujourd'hui ») sont fréquemment corrélés à l'obsolescence. Mais ces tendances ne couvrent pas tous les types de segments d'obsolescence. Pour évaluer ce point et constituer une première méthode naïve de classification, nous testons la prise en compte des traits qui nous semblent les plus probables dans les segments d'obsolescence. Nous comparons ainsi deux systèmes de base s'appuyant sur les traits suivants.

16. http ://www.coheris.fr/fr/page/home.html
17. Le logiciel SPAD fournit une valeur test ou v-test. Les auteurs du logiciel indiquent que si la valeur-test est supérieure à 2, alors le cœfficient est significatif avec un risque d'erreur inférieur à 5 % . Plus la valeur-test est grande (en valeur absolue), plus la liaison entre variables est significative et moins le hasard a de chance d'être responsable de celle-ci.

Base 1 : partant du constat que les traits les plus corrélés à l'obsolescence sont des valeurs numériques et des dates, nous testons tout d'abord si la simple présence de chiffres dans une phrase est suffisante pour déterminer sa nature obsolescente ou non ;

Base 2 : la seconde méthode de base est la présence d'au moins un des traits les plus corrélés à la variable *obsol* : expressions temporelles déictiques, ponctuelles ou de durée lorsqu'elles réfèrent à une date proche du moment d'énonciation ou située dans le futur, les temps/modes futur et conditionnel, les adverbiaux exprimant un point de vue de type récence (« les territoires actuels ») ou prévision (« les recherches à venir »).

	Précision	Rappel	F-score
Base 1	23	31	26
Base 2	30	39	37

Tableau 3. *Performances des systèmes de base*

Les performances sont très mauvaises, tant en précision qu'en rappel avec ce type de systèmes. Ces conclusions nous amènent à développer une méthode rendant compte de l'intérêt de considérer non pas des traits "simples" mais des configurations de traits. C'est pourquoi nous mettons en place une analyse en composantes principales.

5.1.2. *Analyse en composantes principales*

Une analyse en composantes principales (ACP) est une méthode d'analyse multivariée qui permet de représenter des données en utilisant un nombre réduit de dimensions. Ces dimensions, appelées composantes principales, sont basées sur les corrélations des variables initiales, et sont classées par ordre d'importance décroissante.

Nous attendons de cette technique qu'elle fasse émerger les lignes structurantes (*i.e.* les composantes principales) de nos données sur la base de la corrélation entre plusieurs traits et la variable *obsol*. Des travaux comme ceux de (Tanguy *et al.*, 2009) utilisent l'ACP dans un usage similaire.

Si l'on observe les huit premiers axes[18] (ou composantes principales) résultant de notre analyse, des corrélations intéressantes émergent. Par exemple, dans l'axe 2 les variables de type *temporel* et de type *entité nommée géopolitique* sont associées à des variables positionnelles (position finale de phrase) et hiérarchiques et corrélées à la variable *obsol*. L'exemple 7 illustre ce cas[19].

Cet exemple confirme nos intuitions : il s'avère important de repérer les segments qui contiennent des entités nommées et plus spécifiquement lorsqu'il s'agit d'un nom de pays (peu importe sa position), avec une série de valeurs chiffrées qui se rapportent

18. L'ensemble des résultats de l'ACP est reproduit dans (Laignelet, 2009, p. 184-198, 269-294).
19. Il s'agit de l'individu le plus représentatif de l'axe calculé par l'ACP.

Au total, l'**agriculture** n'emploie guère que 1 % des actifs de l'île-de-France, concentrés dans l'**industrie** et de plus en plus dans les **services**, deux secteurs représentés en priorité dans l'**agglomération** de Paris.

n° d'individu : 1237456511712

Exemple 7. *Les entités nommées de type géopolitique (en gras) dans un segment d'obsolescence*

à un domaine géopolitique prédéfini (taux de population, densité, etc.). Dans cette composante, les variables externes (domaine de l'économie, de l'histoire ou de la géographie) sont également corrélés.

Même si nous constatons de nombreux regroupements intéressants (par exemple, les variables de type temporel et de mesure avec la variable $obsol$), l'analyse ne permet pas de réduire efficacement la complexité des données en un nombre réduit de facteurs (le pourcentage de variance des premières composantes reste faible).

D'une manière générale, l'ACP montre que l'obsolescence est un phénomène non trivial et qu'il n'est pas possible de chercher à l'appréhender avec peu de traits ou des configurations de traits simples. Aucune tendance forte n'est exprimée. Pour nous permettre d'automatiser le processus de détection des marqueurs complexes de l'obsolescence, nous mettons en place un système d'apprentissage automatique basé sur des règles d'association.

5.2. *Statistiques prédictives*

Un système d'apprentissage automatique permet d'extraire des régularités à partir d'une masse d'informations. Ces nouvelles connaissances peuvent alors être transposées sur de nouvelles données afin de permettre la meilleure prise de décision possible. Dans notre cas, nous cherchons à décrire l'obsolescence et à formuler des règles (c'est-à-dire des traits/combinaisons de traits) qui permettront par la suite de repérer automatiquement dans des textes nouveaux (c'est-à-dire non annotés manuellement) des segments d'obsolescence.

Les données sont similaires à celles utilisées pour les statistiques de base et l'ACP si ce n'est que nous avons travaillé sur des données discrétisées[20]. De plus, le fait que l'on dispose de la variable $obsol$ (issue de l'annotation manuelle) permet la mise en place d'une technique de type supervisé. Ce système a pour but de valider l'intérêt et la pertinence des traits textuels dans les segments d'obsolescence et de vérifier si une machine peut repérer automatiquement ces segments sur la base de leur repérage.

20. Si une variable est absente sa valeur est égale à 0, si une variable est présente de 1 à n fois, alors sa valeur est égale à 1.

Nous avons choisi d'utiliser un système à base de règles d'association parce qu'il fournit un modèle facilement interprétable (Rioult *et al.*, 2008; Laignelet *et al.*, 2009).

5.2.1. *Un système à base de règles d'association*

Les règles d'association permettent d'extraire de la connaissance sous forme de règles, sans faire d'hypothèse sur le modèle des données, ni choisir *a priori* les descripteurs. Le format des règles est le suivant :

$Si\ A,\ alors\ B$ où :

– A et B sont des conjonctions de traits

– A est la condition, la prémisse de la règle

– B est la conclusion

La vocation de ce type de techniques est exploratoire : toutes les associations possibles sont générées quelle que soit la conclusion produite. Pour notre part, nous nous intéressons aux seules règles concluant sur la classe $obsol$ et s'appliquant au minimum sur 9 phrases.Voici un exemple de règle d'association émergeant de nos données :

$76.895 - exprTemp.nature : déictique; sitTps : coincidence \wedge entiteNom.classe : mesure; sousClasse : evolutif \longrightarrow classe : obsol$

Cette règle stipule que la plupart des phrases qui contiennent une expression temporelle de type *déictique - coïncidence* (« aujourd'hui ») et une entité nommée de type *mesure évolutive* (« 35 hab./km2 ») sont obsolescentes. La mesure indique le score de confiance, fourni par le classifieur, à attribuer à la règle. L'exemple 8 illustre le type de phrases repérées par cette règle.

> Les Noirs, représentent aujourd'hui 12 % de la population ; plus de 50 % d'entre eux sont encore concentrés dans le Sud historique.
>
> Source : Corpus GLI

Exemple 8. *Combinaison de plusieurs traits intraphrastiques*

En ce qui concerne les règles concluant sur la classe $obsol$, elles sont nombreuses (1 500) et redondantes, mais nous ont permis de construire un classifieur automatique.

5.2.2. *Évaluation par rapport aux systèmes de base*

Une évaluation par validation croisée de ce classifieur a été menée : le corpus est découpé en plusieurs parties, dix en l'occurrence ; l'apprentissage se fait sur neuf parties et l'évaluation sur la dixième ; ce processus est réitéré autant de fois qu'il y a de parties.

Cette évaluation fournit les scores de performance suivants indiqués dans le tableau 4 où ils sont comparés aux deux méthodes de base vues précédemment.

	Précision	Rappel	F-Score
Base 1	23	31	26
Base 2	30	39	37
Classifieur	32,9	78,8	46,4

Tableau 4. *Performances du classifieur*

Ces résultats sont encourageants si l'on considère qu'il s'agit là d'une étude exploratoire sur la question de l'obsolescence dans les encyclopédies. Il sont certainement insuffisants pour une exploitation industrielle[21]. Nous avons par ailleurs paramétré notre outil afin que le rappel soit privilégié par rapport à la précision : oublier une révision est plus grave qu'indiquer inutilement un segment à réviser et donc, il vaut mieux repérer trop de segments même s'ils ne sont pas obsolescents.

Pour ce qui est des segments obsolescents que le classifieur ne repère pas, nous avons identifié quelques sources d'erreurs de notre système.

La première cause de silence est directement liée aux erreurs de l'étiquetage de l'outil ALIDIS.

Une seconde cause de silence rappelle le problème évoqué dans la section 2.2 : l'obsolescence concerne une partie de la phrase et non le segment entier, ce qui entraîne un conflit temporel entre les adverbiaux présents dans la phrase. Ainsi, la phrase présentée dans l'exemple 9 n'est pas repérée en tant que segment obsolescent.

> Cette théorie et les principes exposés précédemment sont à la base des modèles cosmologiques élaborés en 1922 par le mathématicien russe A. Friedmann et toujours utilisés aujourd'hui.
>
> Source : Corpus GLI (fiche Histoire - Cortés)

Exemple 9. *Un segment à vérifier qui n'est pas repéré par le classifieur*

Enfin, d'autres sources de silence sont liées à l'absence, dans la phrase, de trait textuel suffisamment fort pour marquer l'obsolescence. Dans certains cas, l'analyse des phrases précédentes et/ou suivantes peut être une piste à affiner.

La section 5.3 met en relief le rôle des traits hiérarchiques, positionnels et externes pour la caractérisation des segments d'obsolescence.

21. La technique des règles d'association n'est sans doute pas la meilleure pour un classifieur. Pour notre objectif de recherche, ce choix nous semble adapté notamment parce qu'il renvoie des résultats interprétables facilement.

5.3. *Mesurer l'impact des différents types de traits textuels*

Pour mesurer et comparer l'impact des différents traits à granularité variable, nous avons créé cinq vues différentes à partir de notre base de données (cf. section 5) :

– *corpusComplet* est la vue décrite jusqu'ici, c'est-à-dire une vue qui prend en compte tous les traits textuels ;

– *corpusIPseuls* est une vue qui prend en compte uniquement les traits intraphrastiques ;

– *corpusIPHierar* est une vue qui prend en compte les traits intraphrastiques et les traits hiérarchiques ;

– *corpusIPPos* est une vue qui prend en compte les traits intraphrastiques et les traits positionnels ;

– *corpusEpure* est une vue « épurée » dans laquelle ne sont conservées que les variables corrélées positivement à la variable $obsol$ (cf. section 5.1.1). Tous les types de traits sont représentés.

Comme cela a été fait précédemment sur le corpus entier, nous avons procédé à un apprentissage automatique à base de règles d'associations.

Dans le tableau 5, les performances liées à chacune de ces vues sont indiquées.

	Précision	Rappel	F-Score
corpusIPseuls	38	37	37,5
corpusIPPos	33,2	56,7	41,9
corpusIPHierar	39,9	45,6	42,5
corpusEpure	38,7	62,3	47,7
corpusComplet	32,9	78,8	46,4

Tableau 5. *Comparaison des performances du classifieur selon les différentes vues sur le corpus d'apprentissage*

Ces résultats montrent tout d'abord qu'exploiter les traits de type intraphrastiques uniquement est insuffisant. Les mesures de précision et de rappel sont très basses (respectivement 38 % et 37 %). En revanche, le gain est mesurable avec l'exploitation de traits de type hiérarchique (+ 5 % sur le F-score), de traits positionnels (+ 4,4 %) ou les deux (+ 8,9 %). De plus, les traits hiérarchiques semblent favoriser la précision alors que les traits positionnels semblent privilégier le rappel. Enfin, il semblerait qu'il vaille mieux ne pas considérer trop de traits textuels : les résultats associés au corpus épuré (qui contient tous les types de traits qui ont été mesurés comme étant significatifs dans les segments d'obsolescence) ont une meilleure précision que lorsque tous les traits sont pris en compte.

6. Des marqueurs de l'obsolescence : le rôle des différents types de traits textuels

Le format des règles d'association renvoie des connaissances nouvelles interprétables sur la base des associations de traits concluant sur la classe *obsol*. Nous proposons un ensemble de marqueurs pour repérer automatiquement l'obsolescence.

Dans les règles d'associations construites par le classifieur, tous les grands types de traits textuels sont représentés :

– des traits hiérarchiques sont associés à des traits intraphrastiques ;

– des traits positionnels textuels sont associés à des traits intraphrastiques ;

– des traits positionnels phrastiques sont associés à des traits intraphrastiques ;

– plusieurs traits intraphrastiques sont associés entre eux ;

– enfin, l'importance du domaine est mise en valeur.

Les marqueurs de l'obsolescence proposés sont ainsi constitués de traits hétérogènes et à granularité variable. Voici quelques exemples de marqueurs.

Tout d'abord, la présence conjointe de plusieurs traits intraphrastiques entraîne la caractérisation obsolescente de la phrase concernée. Plus précisément, le marqueur composé des traits *temporel déictique, coïncidence* et *entité nommée* de type *mesure évolutive* entraîne l'obsolescence du segment. L'exemple 10 illustre ce type de cas.

Aujourd'hui, plus de **25 %** de ces transactions s'effectuent ainsi en euros (contre **48 %** pour le dollar).

Source : Corpus Atlas (fiche Économie - L'euro et les économies européennes)

Exemple 10. *Un segment repéré par un marqueur composé de traits intraphrastiques (en gras)*

Les traits textuels informant sur la position de la phrase dans le paragraphe sont présents dans les règles d'association calculées pour la classe *obsol*. Ainsi, l'association d'une *entité nommée* de type *mesure géopolitique* dans une phrase située en début de paragraphe indique fortement l'obsolescence du segment. L'exemple 11 illustre ce type de marqueur.

§ La population de l'Alsace était estimée à **1 734 145 habitants** au dernier recensement. [...]

Source : Corpus GUL (Géographie - L' Alsace)

Exemple 11. *Un segment repéré par un marqueur composé d'un trait positionnel phrastique (début de paragraphe) et d'un trait intra-phrastique (en gras)*

De la même manière, la position d'introduction de section est très fortement corrélée à des informations *temporelles* ou des entités nommées de type *géopolitique* que ces traits soient intraphrastiques ou hiérarchiques. Ainsi, ce qui semble pertinent pour

les premières phrases de paragraphe l'est également pour les paragraphes introductifs qui ont tendance à situer le décor temporel ou référentiel (*i.e.* de quoi on va parler). Les paragraphes introductifs introduisent souvent un exemple particulier portant sur un phénomène particlier, sur un sujet précis. Il serait intéressant de vérifier si ce sont également des paragraphes qui sont introduits par des marqueurs de discontinuité tels que ceux décrits par (Ho-Dac, 2007).

Quand le titre contient une entité nommée de type *géopolitique* et que la phrase comprend l'association entre des traits *temporels* et des entités nommées de type *mesure*, le segment est obsolescent. La réalisation de ce marqueur est illustrée dans l'exemple 12.

x. **Population**

§ En raison de l'omniprésence à l'ouest du désert du Kalahari, la population se concentre pour l'essentiel dans l'est, le long du grand axe routier et ferré entre l'Afrique du Sud et le Zimbabwe et qui relie les principales villes du pays dont la capitale Gaborone (**195 000 hab.**). [...]

Source : Corpus GUL (Géographie - Le Botswana)

Exemple 12. *Un segment repéré par un marqueur composé d'un trait hiérarchique, d'un trait positionnel (introduction) et d'un trait intra-phrastique (en gras)*

Celui-ci montre le rôle des titres. Ils mettent en place eux aussi ce décor temporel ou référentiel. On observe une forte présence d'entités nommées de type *géopolitique, mesure ou lieu* et d'expressions temporelles de type *déictique* dans les titres prédisant des segments d'obsolescence. On constate également, à la suite de (Ibekwe-SanJuan, 2005) que les connecteurs discursifs et les traits de point de vue sont quasiment inexistants dans les titres. Ces constats sur les titres rejoignent les conclusions de (Ho-Dac *et al.*, 2004) : les titres ont pour fonction de canaliser les connaissances d'arrière-plan (« implication référentielle ») et/ou d'installer ou de mettre le focus sur un référent particulier (« implication thématique »).

Concernant les dernières phrases de paragraphes, elles ont souvent tendance à ouvrir des perspectives sur un phénomène donné et donc à mettre en place des questionnements : le marqueur constitué d'un temps verbal au *conditionnel* avec une entité nommée de type *géopolitique* dans la dernière phrase du paragraphe entrainera l'obsolescence du segment. Les marqueurs composés du trait positionnel de dernière phrase de paragraphe sont moins productifs mais ont été malgré tout générés : ainsi la configuration d'un argumentatif de type *correction* (« mais ») avec une entité nommée de type *géopolitique* dans une phrase en fin de paragraphe entraînera la classification du segment comme obsolescent.

La position conclusive est, elle aussi, associée aux entités nommées de type *mesure* dans les règles d'association : toujours en parallèle avec ce que nous avons dit précédemment sur les dernières phrases de paragraphes, un paragraphe conclusif introduit des exemples précis, concrets qui font appel à des valeurs chiffrées précises et qui seront donc plus souvent soumises à ré-évaluation.

Enfin, la caractérisation du texte en fonction des rubriques thématiques est également un trait important pour la recherche des segments d'obsolescence et fait partie intégrante des marqueurs de l'obsolescne constitués. Certaines combinaisons de traits ne sont pertinentes que pour une rubrique particulière. Par exemple, le marqueur combinant les traits *entité nommée lieu* et *entité nommée mesure* est fortement associée à l'obsolescence dans un texte géographique mais il est contre-productif en histoire.

7. Conclusion et perspectives

D'une manière générale, le phénomène de l'obsolescence semble une réponse appropriée à la question de la mise à jour dans les documents encyclopédiques. Nous avons montré que les segments d'obsolescence peuvent être caractérisés par des configurations de traits textuels dont une importante particularité est d'être de granularité variable.

La notion de marqueur discursif, qui suppose la prise en compte de combinaisons de traits textuels hétérogènes, est centrale dans notre approche. Nous avons notamment montré que l'ensemble des types de traits textuels pris en compte est nécessaire puisque toutes les associations possibles sont représentées dans les règles d'association concluant sur la classe *obsolescence*. Par ailleurs, l'importance de la structure du discours a été montrée, essentiellement à travers le rôle joué par les titres mais également la position des segments et l'importance du domaine.

L'observation approfondie des segments à mettre à jour qui ne sont pas repérés par notre classifieur est centrale pour améliorer l'outil et rendre le système plus performant. Nous avons notamment évoqué la nécessité d'améliorer les repérages des traits textuels faits par l'outil ALIDIS mais également le problème de la délimitation des segments (lorsqu'ils sont inférieurs aux frontières phrastiques) ou encore la question du découpage temporel.

La méthodologie mise en œuvre est évolutive (tant les programmes que les ressources), reproductible en partie (sous réserve des corpus sous licence) et réutilisable. Nous souhaitons poursuivre ces travaux à la fois en exploitant d'autres corpus de type encyclopédique mais aussi en affinant la notion d'obsolescence, notamment à travers la distinction de deux types de segments d'obsolescence : les segments à réadapter (information évolutive) et les segments à réactualiser (information plus actuelle, plus pertinente). L'idée que certains traits et configurations de traits puissent être pertinents selon ces types nous intéresse particulièrement.

Le modèle présenté dans cet article exploite 150 traits textuels différents. Il y aurait certainement lieu d'affiner ce panel de traits textuels, par élimination et/ou fusion pour les traits évalués, et par ajout de nouveaux traits en fonction d'autres regards linguistiques.

Nous souhaitons également travailler sur des phénomènes différents de celui de l'obsolescence afin de mettre notre méthodologie à l'épreuve. Par exemple, il serait

intéressant d'éprouver le rôle des traits hiérarchiques et positionnels sur l'étude de segments argumentatifs (cf. (Teufel, 1999)) ou encore sur l'étude des énumérations (cf. (Péry-Woodley, 1994)). Des travaux ayant déjà été menés sur ces thématiques, cela permettrait entre autres de mesurer les apports de notre méthode.

Remerciements

Nous tenons à remercier François Rioult, à qui la section 5 de ce travail doit beaucoup, et les trois relecteurs anonymes pour leurs commentaires judicieux et constructifs.

8. Bibliographie

Asher N., Lascarides A., « Temporal Interpretation, Discourse Relations, and Commonsense Entailment », *Linguistics and Philosophy*, vol. 16, p. 437-493, 1993.

Biber D., *Variation across speech and writing*, Cambridge University Press, Cambridge, 1988.

Biber D., Connor U., Albin-Upton T., *Discourse on the move : using corpus analysis to describe discourse structure*, John Benjamins, 2007.

Bilhaut F., Analyse automatique de structures thématiques discursives - Application à la recherche d'information, Thèse de doctorat, Université de Caen, 2006.

Borillo A., « Aide à l'identification des prépositions composées de temps et de lieu », *Lexique*, vol. 11, p. 175-184, 1997.

Bouffier A., Analyse discursive automatique de textes - Application à la modélisation de textes incitatifs, Thèse de doctorat, Université Paris Nord - Villetaneuse, 2008.

Charolles M., « Cohésion, cohérence et pertinence du discours », *Travaux de Linguistique*, 1995.

Charolles M., « L'Encadrement du Discours, Univers, Champs, Domaine et Espaces », *Cahiers de Recherche linguistique*, 1997.

Desclés J.-P., Guentcheva Z., Maire-Reppert D., Oh H.-G., « A propos de la catégorie grammaticale du temps et de l'aspect », *Parcours Linguistique de discours spécialisés*, Editions Scientifiques Européennes, Peter Lang S.A., Paris, p. 291-299, 1992.

Gosselin L., *Temporalité et modalité*, de Boeck-Duculot, 2005.

Ho-Dac L.-M., Jacques M.-P., Rebeyrolle J., *Sur la fonction discursive des titres*, Pleyben, p. 125-152, 2004.

Ho-Dac L.-M., Péry-Woodley M.-P., Tanguy L., « Anatomie des structures énumératives », *Actes de TALN 2010*, Montréal (Québec), 2010.

Ho-Dac M., La position initiale dans l'organisation du discours : une exploration en corpus, Thèse de doctorat, Université de Toulouse 2 - Le Mirail, 2007.

Ibekwe-SanJuan F., « Annotation d'indices de nouveautés dans les écrits scientifiques et techniques », *Colloque Indice, Index, Indexation*, 2005.

Jacques M.-P., Rebeyrolle J., « Titres et structuration des documents », *Actes du Colloque International Discours et Document*, Caen, France, p. 1-12, 2006.

Laignelet M., Analyse discursive pour le repérage automatique de segments obsolescents dans les documents encyclopédiques, Thèse de doctorat, Université de Toulouse - Le Mirail, 2009.

Laignelet M., Rioult F., « Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs », *Actes de TALN 2009*, 2009. Prix du Meilleur Papier.

Laurendeau P., « Modalité ; opération de modalisation et mode médiatif », *in* P. de l'université de Rouen (ed.), *Les médiations langagières - Des faits de langue aux discours*, vol. 1, R. Delamotte-Legrand, 2004.

Mani I., *Automatic summarization*, John Benjamins Publishing Compagny, Amsterdam/Philadelphie, 2001.

Péry-Woodley M.-P., « Une pragmatique à fleur de texte : marques superficielles des opérations de mise en texte », *Parcours Linguistique de discours spécialisés*, Editions Scientifiques Européennes, Peter Lang S.A., Paris, p. 337-348, 1994.

Péry-Woodley M.-P., *Sémantique et Corpus*, Hermès - Lavoisier, Paris, chapter Discours, corpus, traitements automatiques, p. 177-210, 2005.

Power R., Scott D., Bouayad-Agha N., « Document Structure », *Computational Linguistics*, vol. 29, n° 2, p. 211-260, 2003.

Rioult F., Zanuttini B., Crémilleux B., « Apport de la négation pour la classification supervisée à l'aide d'associations », *Conférence d'Apprentissage*, p. 183-196, 2008.

Tanguy L., Tulechki N., « Sentence Complexity in French : a Corpus-based Approach », *proceedings of the 17th conference on Intelligent Information Systems (IIS)*, Krakow, 2009.

Teufel S., Argumentative Zoning, PhD thesis, Université de Edimbourg, 1999.

Weinrich H., *Le temps*, Editions du Seuil, Paris, 1973.

Widlöcher A., Bilhaut F., « La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus », *Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN)*, Dourdan, France, 2005.

Zerida N., Lucas N., Crémilleux B., « Combinaison de descripteurs linguistiques et de structure pour la fouille d'articles biomédicaux », *Schedae*, p. 69-78, 2006.